Introducing an Annotated Bibliography on Temporal and Evolution Aspects in the World Wide Web

Fabio Grandi
IEIIT.BO-CNR and DEIS
Alma Mater Studiorum – Università di Bologna, ITALY
fgrandi@deis.unibo.it

1. INTRODUCTION

Time is a pervasive dimension of reality as everything evolves as time elapses. Information systems and applications at least mirror, and often have to capture, the time-varying and evolutionary nature of the phenomena they model and the activities they support. This aspect has been acknowledged and long studied in the field of temporal databases but it truly applies also to the World Wide Web, although it has not seemingly considered as a primary issue yet. However, several papers addressing, in an explicit or implicit way, the representation and management of time and change in the World Wide Web appeared recently and, on some aspects, showed a clear upward trend in last months, witnessing a sustained and/or growing interest.

Reflecting and acknowledging such interest, we started in 2001 to collect references concerning the handling of time and evolution issues in World Wide Web research. The purpose of this collection was to compile a bibliography which could be of help, in particular, to students and young researchers. As a result of such endless work, we wrote an annotated bibliography, whose latest version is available on the Web at URL:

`http://www-db.deis.unibo.it/~fgrandi/TWbib/`

This follows several fortunate bibliographies on time-varying information, including seven ones on temporal databases (Bolour, Anderson, Dekeyser & Wong, *ACM SIGMOD Record* **12**:3, 1983; McKenzie, *ACM SIGMOD Record* **15**:4, 1986; Stam & Snodgrass, *IEEE Database Engineering* **17**:4, 1988; Soo, *ACM SIGMOD Record* **20**:1, 1991; Kline, *ACM SIGMOD Record* **22**:4, 1993; Tsotras & Kumar, *ACM SIGMOD Record* **25**:1, 1996; Wu, Jajodia & Sean Wang, in *Temporal Databases: Research and Practice*, Springer LNCS 1399, 1998), two ones on spatio-temporal databases (Al-Taha, Snodgrass & Soo, *ACM SIGMOD Record* **22**:1, 1993; Al-Taha, Snodgrass & Soo, *Intl' Journal of Geographical Inf. Sys.* **8**:1, 1994), two ones on spatio-temporal data mining (Roddick & Spiliopoulou, *ACM SIGKDD Explorations* **1**:1,1999; Roddick, Hornsby & Spiliopoulou, *Proc. TSDM 2000 Intl' Workshop*, Springer LNCS 2007, 2000), one on schema evolution (Roddick, *ACM SIGMOD Record* **21**:4, 1992), and one on (temporal) indeterminacy (Dyreson, in *Uncertainty Management in Information Systems*, Kluwer, 1996).

In the following, we introduce the bibliography contents by reproducing here, with the same section organization of the paper, the annotations that have been included. This can be used as a quick reference for locating the section(s) which contain the works of interest in the full bibliography.

2. PAPERS

The references, which amount to nearly 600 as of September 2003, are organized into 10 main sections, where they are further grouped by some similarity criterion introduced by brief notes. The collected bibliographic entries include published papers and Web resources but also theses and technical reports (if unpublished or significantly containing —as far as we know— more material than their published version). Entries are listed in ascending chronological order within (sub)groups. URLs have been checked as of July through August 2003.

2.1 Transaction Time

Transaction time is the time some fact is current in a database, from when it is stored in the system to when it is deleted. With respect to the Web, it represents the on-line availability and versioning of resources in a Web site, even if they are not created by "transactions" in a strict sense. Although in several cited papers it was called "valid time", if it concerns the time Web documents are modified it is properly transaction time. The number of references in this section is 29.

The first group of papers explicitly deals with transaction-time versioning of Web resources, which can be selectively accessed, on user's request, through a transaction-time aware temporal Web server.

The second group of papers deals with temporal extensions of data models (e.g. OEM and XML) and query languages for semistructured data. Since data versions are generated by updates (and without an explicit management of time), the involved temporal dimension is obviously transaction time.

2.2 Valid Time

Valid time is the time some fact is true in the real world. In the Web, it concerns the temporal validity of the information carried by the contents of a Web resource. The number of references in this section is 71.

In the first group of papers, modeling and management of historical Web documents is properly concerned.

Several papers (e.g. in the humanities field), though not dealing with temporal data modeling and querying in a strict sense, consider (valid) time as a useful annotation dimension. Some works are also mainly concerned with extraction and visualization of temporal information available on the Web.

A strictly related field concerns the addition of the (validity) temporal dimension to virtual environments (e.g. temporal VRML extensions).

2.3 Bitemporal (Transaction and Valid Time)

In the temporal database literature, bitemporal means transaction plus valid time data. Some papers address both time dimensions also in the World Wide Web context. Additional time dimensions (e.g. user-defined times) are also considered in some works. The number of references in this section is 18.

2.4 Non-Temporal Versioning

Consistent research efforts have been devoted to the management of (non-temporal) resource *versions* over the Web. The number of references in this section is 66.

A first stream of research, where the main focus is on collaborative work and distributed document authoring, is concerned with maintenance of versions as produced by sequences of modifications. The implicit temporal dimension underlying the editing process (i.e. transaction time) is neglected.

A second stream of non-temporal versioning research involves the design and implementation of (multidimensional) *context-dependent* Web resources, supporting for example personalization, location-based services, multi-device and multi-lingual presentations, etc. Since this is a large bibliography, we tried to collect references most relevant to data-intensive applications (or also considering temporal aspects).

Other papers on adaptive techniques for Web personalization based on user profiling can also be found at the end of **Sec. 2.7**.

2.5 Updates on the Web

Several recent works are focused on the management of updates of Web resources like XML repositories. No time dimension nor versioning is usually considered. The number of references in this section is 102.

Several works specifically deal with the detection (and measurement) of changes in Internet/Web resources.

Some works also considered modification or evolution of the *schema* of semistructured data.

2.6 Time in Web Warehousing

Several papers, including many from the previous section, deal with the management of (virtual) Web warehouses or (materialized) views over semistructured data. Temporal aspects, either related to data modifications or to information contents, are often considered. These range from the (valid) time dimension involved in multidimensional data modeling and OLAP, to maintenance of temporal consistency of the information in the warehouse. Similar aspects are also involved in the maintenance of dynamically generated Web pages. The number of references in this section is 47.

Temporal aspects, detection and management of changes, updating and versioning of Web resources have also been considered together in the context of comprehensive projects involving repositories/warehouses of web sources (actually, also some of the papers in the previous section deal with more than one such aspects; for the sake of classification, their taxonomy is based on what has been judged the main focus, though sometimes somehow arbitrarily). These projects are specifically introduced and described in some references.

2.7 Navigation Time

An autonomous research thread concerns time in Web usage. We can define *navigation time* the temporal dimension marking the navigation of the Web by a user. A user's navigation history also corresponds to what has been called "clickstream", and analysis of Web logs (e.g. for Web user profiling) is also known as Web usage mining. The number of references in this crowded section is 181.

The "history" facility of Web browsers is based on rollback along navigation time. Several papers focused on equipping browsers with enhanced history mechanisms.

The temporal dimension explicitly used for historical queries in Web warehouses concerns the time the warehouse is built. Hence, it corresponds to the navigation time with respect to the source Web sites (but also to the "transaction" time in the warehousing system).

Furthermore, also the various Internet archiving initiatives have to be referenced here. They consist in building large archival libraries of Web pages, by storing periodic backup copies of Web sites. Such an archive can be provided with a temporal search engine and, in practice, represents a Web warehouse with support of transaction time (which corresponds to the navigation time the Web pages were accessed to be copied).

Another temporal aspect in Web navigation involves the human-computer interaction (e.g. Web response time and its perception by users). Moreover, a strictly related issue is the "optimization" of such interaction leading, for example, to the development of caching and prefetching techniques or to site restructuring (e.g. personalization). Such issue is covered in a quite broad specific literature. We tried to reference most of the papers where the proposed techniques are, in turn, based on temporal or evolutionary aspects (e.g. analysis of navigation histories).

2.8 Time and Change in Semantic Web

Temporal and evolution aspects have also been considered in the definition of *ontologies* for the so-called *semantic Web*, and further developments are likely to appear in the next future. Furthermore, some papers specifically deal with evolution and versioning of ontologies. 16 references are listed in this section.

2.9 WWW Evolution

Several papers deal with the history and evolution of the Web itself. Technical issues but also "cultural" (e.g. sociological or media) aspects have been considered. This is also a broad literature, for which a separate bibliography could be compiled. For the sake of brevity, we have only cited a few representatives.

Some papers deal with the evolution of the Web development and usage.

Other papers deal with the evolution of single Web sites (e.g. software engineering case studies) or of Web site and service types.

The total number of references in this section is 30.

2.10 Other Time Aspects

A miscellanea of other 28 citations can eventually be found in this section.

Several efforts involve standardization of temporal expressions to be used on the Web (e.g. as metadata). For instance, the issue has been addressed by the W3C Consortium, and is part of the XML-Schema datatype definitions. Metadata carrying temporal information have also been considered in the TEI, Dublin Core, MUC, EAD and Open Archives initiatives.

Some multimedia Web resources have an intrinsic temporal dimension connected with the user's fruition of the resource itself. These include streaming video and music, animated presentations and so on. For the sake of completeness, we only listed a few representative references; although dealing with time, they quite represent a world apart.

Finally, according to some authors, the Web itself has developed its own time dimensions (e.g. "Internet time", "virtual time", etc).

3. FINAL REMARKS

Additions, corrections and comments concerning the bibliography are obviously welcome. Anyway, we apologize (with the readers and especially with the authors) for any **errors**, **misclassifications** and **omissions** may result from the collected entries.

Last but not least, the bibliography could not have been compiled (i.e. in reasonable time) without the existence and availability of the World Wide Web and, in particular, without the Google search engine, the ACM Digital Library and the IEEE Xplore service, and all the priceless information supplied by Michael Ley's DBLP Computer Science Bibliography and the NEC CiteSeer Research Index.